The Writing Pal Intelligent Tutoring System: Usability Testing and Development

Rod D. Roscoe a,*, Laura K. Allen b, Jennifer L. Weston b, Scott A. Crossley c, Danielle S. McNamara d

a Human Systems Engineering in the Ira A. Fulton Schools of Engineering at Arizona State University, United States
b Psychology and the Learning Sciences Institute at Arizona State University, United States
c Applied Linguistics and ESL at Georgia State University, United States
d Learning Sciences Institute and Professor of Psychology at Arizona State University, United States

Abstract

The Writing Pal (W-Pal) is an intelligent tutoring system (ITS) designed to improve students’ writing proficiency via a unique combination of explicit strategy instruction, game-based practice, essay writing practice, and automated formative feedback. To develop and refine the many features of the W-Pal tutoring system, we have employed a multiple-method usability testing approach, which capitalizes on the complementary strengths and weakness of methods such as focus groups, component studies, internal testing, and in vivo testing. These diverse methods allow researchers to benefit from focused student input, instructor input, and iterative development, while also gathering data in ecologically-valid settings. In this paper, we describe some of the testing and development of aspects of W-Pal, consider the challenges of building such a system, and provide a particular emphasis on a feasibility study that integrated W-Pal into high school English classrooms during a school year. The results of this study showed that students perceived the system as informative, valuable, and enjoyable, and results also highlighted specific ways that these aspects of the system could be further enhanced. Based upon these findings, a significantly updated version of W-Pal has been developed. The current W-Pal system is described along with considerations for future research and how the system may be used to supplement writing instruction.

© 2014 Elsevier Inc. All rights reserved.

Keywords: Writing instruction; Writing strategies; Intelligent tutoring systems; Automated writing evaluation; Formative feedback; Educational games; Usability testing

1. The Writing Pal intelligent tutoring system: Usability testing and development

The Writing Pal (W-Pal) is an intelligent tutoring system (ITS) designed to improve students’ writing proficiency via a unique combination of strategy instruction, game-based practice, essay writing practice, and automated formative feedback. This novel technology has been developed over several years of iterative, interdisciplinary work that synthesizes best practices from writing pedagogy and user-centered design (Dai, Raine, Roscoe, Cai, & McNamara, 2011; McNamara et al., 2012), with contributions from psychology, composition, linguistics, and computer science.

* Corresponding author.
E-mail address: rod.roscoe@asu.edu (R.D. Roscoe).

http://dx.doi.org/10.1016/j.compcom.2014.09.002
8755-4615/© 2014 Elsevier Inc. All rights reserved.
This work incorporates design principles common to many ITSs, such as cross-platform accessibility (i.e., PC and Mac) at school and at home, and database structures that enable robust data logging and continuation of service despite interruptions.

W-Pal has also been informed by fundamental pedagogical design principles. First, W-Pal was designed with the assumption that writing is a complex process with several non-linear phases, such as prewriting, drafting, and revising (e.g., Deane et al., 2008; Flower & Hayes, 1981). Numerous strategies can be taught to support these phases (e.g., Graham & Perin, 2007; Hillocks, 1984), and teachers need freedom to utilize W-Pal in ways that meet diverse curricula. Thus, W-Pal presents strategies via multiple Strategy Modules that can be studied and practiced in any order. Second, the development of writing proficiency requires opportunities for sustained practice, yet student engagement with ITSs can decrease over time, leading to disengagement (Bell & McNamara, 2007). To ameliorate these effects, W-Pal offers numerous educational games that enable strategy practice, which are hypothesized to foster motivation (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005; McNamara, Jackson, & Graesser, 2010). Third, writing development requires individualized and formative feedback (McGarrell & Verbeem, 2007; Sommers, 1982). This is a significant challenge because writing and assessment are highly subjective (Huot, 1996; Meadows & Billington, 2005). W-Pal, like human instructors, must assess open-ended input, including essays and writing excerpts, and provide meaningful responses. For this purpose, we have compiled a number of natural language processing (NLP) tools to inform our writing assessment algorithms. These tools include a lemmatizer, syntactic parsers, lexical databases (Princeton University’s WordNet, the Max Planck Institute for Psycholinguistics’s CELEX, and the MRC Psycholinguistic Database), rhetorical analyzers, and Latent Semantic Analysis (LSA; Graesser & McNamara, 2011; McNamara & Graesser, 2012). NLP algorithms also drive feedback in W-Pal. This, too, has been challenging because there is little empirical research on how to design and implement formative feedback based on automated assessments (Grimes & Warschauer, 2010; Roscoe, Kugler, Crossley, Weston, & McNamara, 2012; Roscoe et al., 2011; Roscoe, Varner, Crossley, & McNamara, 2013).

W-Pal is not alone in providing automated writing assessment and feedback. A number of automated writing evaluation systems (AWEs) are available. These programs score essays and offer feedback via a combination of statistical modeling, NLP, LSA, machine learning, and other artificial intelligence methods. Systems such as Education Testing Service’s Criterion (Attali & Burstein, 2006; Burstein, Chodorow, & Leacock, 2004) and Vantage Learning’s MY Access! (Grimes & Warschauer, 2010; Rudner, Garcia, & Welch, 2006) rely primarily on statistical modeling. In such systems, an essay corpus is annotated to identify essay elements (e.g., topic sentences). Essays are then analyzed along many linguistic dimensions, and statistical analyses extract features that discriminate between higher- and lower-quality essays. Statistical models combine extracted linguistic properties into algorithms that assign grades to essays. In contrast, the Pearson Intelligent Essay Assessor (Landauer, Latham, & Foltz, 2003; Wohlpart, Lindsey, & Rademacher, 2008), uses LSA to assess essays. LSA assumes that similar words occur in similar contexts (e.g., sentences, paragraphs, or whole texts). Singular value decomposition is used to reduce a large word-by-context matrix to the number of dimensions needed to capture semantic structure. Essay scores are based on the semantic similarity between student essays and a benchmark corpus of pre-scored essays.

Proponents of AWEs argue students receive more opportunities to write and receive feedback without adding to instructors’ workload. Providing timely, personalized feedback on student writing is rewarding but time-intensive, and AWEs allow teachers to assign more writing than might otherwise be feasible. However, two objections to AWEs are that they lack human sensitivity and that the classification of writing quality is limited by available algorithms (Hearst, 2002). AWEs, with their reliance on broad statistical regularities, may not capture writers’ unique individual expressive differences (Clauser, Kane, & Swanson, 2002). In addition, savvy users can potentially trick the scoring system (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). Thus, despite progress, automated scoring systems are still under development with areas for improvement. Further information about the history and development of such scoring tools can be found in overviews by Mark Warschauer and Paige Ware (2006) and Semire Dikli (2006).

The fundamental differences between W-Pal and AWEs reside in their respective origins and pedagogy. AWEs such as MY Access! began as scoring systems (e.g., the core of MY Access! is the Vantage Learning IntelliMetric scoring engine). The purpose of such tools was to facilitate the rating of essay quality. Indeed, much of the AWE literature focuses on demonstrating scoring reliability and accuracy (Grimes & Warschauer, 2010). Over time, algorithms increased in sophistication and could address finer-grained essay traits that could then be communicated to students as feedback. In short, the conceptualization of AWEs began with scoring and assessment but has seen a gradual accumulation of pedagogical utility.
W-Pal’s conceptualization has proceeded along a nearly opposite trajectory. From its inception, W-Pal has been focused on writing instruction and writing strategies (i.e., pedagogy). Automated scoring is a necessary component of W-Pal, but it is not its defining feature. Rather than providing feedback only upon what is most easily or reliably scored, we strive to develop algorithms that can detect the use of writing strategies (or lack thereof) employed by developing writers. This is a difficult task, but one that may afford feedback that is more valid or useful to students. In short, W-Pal shares some overlap with AWEs, but is designed to serve a distinct pedagogical role in composition instruction. Importantly, W-Pal is not intended to replace expert and experienced teachers. W-Pal is intended to supplement classroom teaching by offering more opportunities for writing strategy development, practice, and feedback.

In summary, W-Pal is a complex ITS designed to supplement writing instruction by reinforcing instructors’ strategy instruction and offering more opportunities for practice and feedback. To develop and refine W-Pal, we have employed a multiple-method usability testing approach (e.g., Holzinger, 2005; Proctor & Vu, 2007). W-Pal development has benefited from canonical usability testing methods, such as focus groups and component experiments, which assess end-users’ interactions with system features and inform incremental improvements in functionality and performance. Such testing thus embodies two key principles: recruitment of end-users and rapid iterations. W-Pal development has also gained from alternative approaches that selectively violate these principles. In internal usability testing (akin to a cognitive walkthrough), the researchers take on the role of the end-users and attempt to use the system as designed. Finally, in vivo usability testing sacrifices the speed of rapid iterations to explore the holistic system in a natural setting over time. In this paper, we showcase example studies from this multiple-method approach. Based on these data, we consider the possibilities and constraints offered by W-Pal for writing instruction, research, and assessment. We then discuss how W-Pal might be integrated into classrooms and discuss related questions for future research. Importantly, our goal is not to establish the efficacy of W-Pal, but to establish its feasibility and discuss key issues in the design of such a technology.

2. Development of Writing Pal: Version 1

2.1. Teacher focus groups to inform early content and design

The strategies taught in W-Pal are grounded in writing instruction research and have been refined via input from expert classroom educators. Exemplifying this research base, Steve Graham and Dolores Perin (2007) reported in their meta-analysis of over 120 interventions that the most effective programs were those that explicitly taught students strategies for planning, drafting, editing, and summarizing. This analysis quantified the impact of interventions using the Cohen’s $d$ statistic, a measure of the magnitude of a statistical effect size (Cohen, 1969). As a rule of thumb, effect sizes around .20 are considered small, effects around .50 are moderate, and effects above .80 are large. Strategy-based interventions demonstrated an effect size of $d = .82$ versus comparison groups; strategy-based interventions promoted better performance than programs without strategy training. Formats such as peer learning ($d = .75$), goal-setting ($d = .70$), and prewriting ($d = .32$) also contributed positively. The one exception was grammar and spelling instruction ($d = -.32$), which contributed negatively. An exemplar of this research is provided by Susan de la Paz and Steve Graham (2002), who taught adolescents to use a PLAN (i.e., Pay attention to the prompt, List main ideas, Add supporting ideas, Number your ideas) and WRITE (i.e., Work from your plan to develop your thesis statement, Remember your goals, Include transition words for each paragraph, Try to use different kinds of sentences, and Exciting, interesting, $100,000$ words) strategy. The PLAN mnemonic helped students attend to the prompt, generate main and supporting ideas, and organize these ideas. The WRITE mnemonic helped students use their plans during writing, and to use varied sentence structure and vocabulary. A control condition received traditional writing instruction, which included lessons on grammar, spelling, vocabulary, generating ideas, and organization. Overall, students trained to use PLAN and WRITE generated more and better plans ($d = 1.17$), wrote longer essays ($d = .82$), used more sophisticated vocabulary ($d = 1.13$), and wrote better essays ($d = 1.71$) at posttest.

Initial W-Pal prototypes, informed by strategy research and writing pedagogy, were presented to two focus groups of high school English educators (Kim et al., 2012): a group of eight teachers from Memphis, TN and a second group of six teachers from the Washington, DC area. As is commonly encountered in focus groups, challenges arose due to social pressures wherein outspoken participants could sway the apparent views of the others. To overcome this limitation, potential responses and ideas were collected periodically and transformed into brief questionnaires that allowed participants to respond freely and individually. Participating teachers were also interviewed individually. This
iterative and interactive process revealed many potential applications for W-Pal, such as group and individual classwork, homework, and tutoring. Teachers desired a flexible system that was sensitive to differences in student ability, classroom sizes, and technology resources. Teachers also emphasized grade-level appropriate wording and examples, and favored a peer-based rather than authoritative approach. These focus groups generated a comprehensive list of goals and strategies for 1) writing introductions and conclusions, 2) expressing ideas in a unique voice, 3) understanding the prompt, 4) stating valid evidence, 5) elaborating essays, 6) writing thesis statements, 7) paraphrasing, 8) avoiding choppy or run-on sentences, 9) structuring the essay, 10) generating ideas, 11) organizing ideas, 12) writing topic sentences, and 13) improving essay cohesion.

Teachers’ input informed the development of eight writing Strategy Modules (and one introductory Prologue module) that taught strategies for prewriting (Freewriting and Planning), drafting (Introduction Building, Body Building, and Conclusion Building), and revising (Paraphrasing, Cohesion Building, and Revising). Each module contained a 20–30 minute lesson video in which animated pedagogical agents—a teacher agent named Dr. Julie and two student agents, Mike and Sheila—discussed the strategies. The student characters allowed learners to have peer-like models for instruction. Animated characters were created using MediaSemantics Character Builder with text-to-speech (TTS) voices by Loquendo, which provide high flexibility for character gestures and voices. Writing strategies were presented with many examples and illustrations and were often organized by mnemonic devices.

Mnemonic devices were created by our team and modeled after successful mnemonics in prior research (e.g., de la Paz & Graham, 2002), which made the goals and criteria of writing strategies explicit and memorable. For example, the Freewriting module FAST PACE mnemonic summarized different ways to quickly freewrite and generate ideas (i.e., Find evidence, Ask and answer questions, Spell out your argument, Think about both sides, Prompt: reread it, Add details, Connect ideas, and Examples: add more). Similarly, the Conclusion Building RECAP mnemonic reminded students that conclusions should provide a “recap” summary of their essay and meet particular goals (i.e., Restate the thesis, Explain how arguments supported the thesis, Close the essay, Avoid adding new evidence, and Present the conclusion in an interesting way).

2.2. Usability Experiments to Guide Module Development: Freewriting Example

Usability experiments were used to test and further develop individual modules. For example, the Freewriting module used the FAST PACE mnemonic to provide students with techniques to generate ideas prior to writing (e.g., Hilgers, 1980). One question that arose early in development was how to structure freewriting practice. With regards to interface design, would students freewrite more effectively if the text box was larger or smaller? A larger space could encourage more writing but might overwhelm some learners. Similarly, what are the effects of repeated freewriting practice? Would students improve by practicing more frequently or would fatigue effects inhibit gains? Finally, how would the W-Pal lesson compare to similar lessons compiled from existing sources? These concerns were also relevant to how information and practice could be offered in the other W-Pal strategy modules.

These questions were investigated in an experiment with 51 college students (Raine, Mintz, Crossley, Dai, & McNamara, 2011). Participants were taught to freewrite by viewing the W-Pal lesson or a control condition video. The control video was compiled from freely available resources online, which discussed style and formatting, plagiarism, referencing, and freewriting. Subsequently, participants wrote six practice freewrites using one of three text-box sizes (1, 10, or 20 lines), with two freewrites per box size. Analyses indicated that freewrite quality did not differ based on the video viewed or box size. Although this study provided disappointing results for W-Pal, it also revealed design issues that may have contributed to the observed outcome. Specifically, freewrite quality declined during practice due to fatigue and boredom. Thus, the potential value of W-Pal was undermined by a presentation style that did not engage the learners.

Based upon the results of such studies, Version 1 strategy modules were redesigned to incorporate more interactivity and engaging practice. First, usability testing results confirmed the need for including lesson Checkpoints to reinforce concepts and hold students’ attention. In each lesson, we embedded additional activities that reinforced the content via short quiz-like or game-like tasks. Checkpoints offered recognition (e.g., identify examples), organization (e.g., manipulate text), and/or generation (e.g., write text) practice. Second, findings such as the decline in freewrite quality due to fatigue highlighted a need for engaging forms of practice. Thus, in addition to changes to the lessons, each module was enhanced with one or more educational games (Dai, Raine, Roscoe, Cai, & McNamara, 2011; McNamara et al., 2012). Educational games can improve students’ motivation by leveraging their intrinsic enjoyment of gaming.
(Orbach, 1979; Schank & Neaman, 2001), and game elements (e.g., competition) can be integrated in ITSs to create a more enjoyable experience (McNamara, Jackson, & Graesser, 2010). Games also require the application of skills and knowledge to succeed or explore (Gredler, 2004), which is similar to computer-based learning environments that require the application of skills and knowledge to complete learning tasks. Thus, games may offer students additional motives to engage in the learning tasks.

To further exemplify how strategy modules were enhanced, Figure 1 provides an example from Introduction Building. In this module, students learn about essay introductions (e.g., Henry & Roseberry, 1999; Hyland, 1990) and relevant strategies. These strategies were summarized by the TAG mnemonic: Thesis statements, Argument preview, and Grab the reader’s attention. Several attention-grabbing techniques were discussed, such as asking thought-provoking questions, which were based on common recommendations from writing style guides (e.g., Hacker, 2009). Embedded checkpoint activities included quizzes and game-like tasks. For example, in the Mission to the Moon mini-game, players earned “moon rocks” by identifying attention-grabbing techniques used in an example introduction. Completing the lesson unlocked several educational games, such as Essay Launcher (Figure 2). In this game, players repaired spaceships by identifying appropriate thesis statements for an example introduction. To navigate, students indicated which attention-grabbing technique was exhibited in the paragraph. Incorrect answers triggered a malfunction animation and feedback, whereas correct answers rewarded students with an animation showing the ship returning to Earth.

2.3. Internal usability testing to evaluate essay scoring and feedback

Another essential component of W-Pal was the Essay Writing Interface, which enabled students to write essays and receive automated formative feedback. Students could select a prompt, set a time limit, and use a scratchpad for planning. Students wrote using a simple word processor and submitted their finished essays to W-Pal for feedback (Figure 3). Underlying the feedback system were a variety of algorithms that assessed the essays and guided the feedback delivered to the students. These algorithms were informed by prior computational linguistics research investigating the links between essay quality and linguistic, semantic, rhetorical, and syntactic text properties (Crossley & McNamara, 2011; McNamara, Crossley, & McCarthy, 2010). These algorithms combined NLP, LSA, machine learning, and other methods.
Figure 2. Essay Launcher allows students to practice thesis statement and attention-grabbing strategies.

All essays received a holistic rating from “Poor” to “Great” (a six-point scale similar to the SAT scoring rubric). Each essay also passed through a series of thresholds to assess Legitimacy (e.g., nonwords), Length, Topic Relevance, and Structure. Essays received feedback only at the highest level of threshold failed. Thus, an essay that passed Legitimacy and Length, but was off-topic, would only receive feedback on Relevance. Essays that that passed all initial

Figure 3. The essay Feedback Report provides a holistic rating and strategy recommendations for revising.
checks would then be assessed based on the quality of introduction, body, and conclusion paragraphs (e.g., Roscoe). If any essays’ sections fell below certain criteria, students received additional feedback on drafting those sections. This feedback also incorporated general revising strategies.

Prior to deployment with students, it was important for us to evaluate feedback usability. Rod Roscoe et al. (2011) collected a corpus of 201 essays written by college students, which had been previously scored by expert raters. These essays were submitted to the W-Pal essay feedback system. Two members of the research team then took on the role of student users and attempted to revise a subset of 48 essays (chosen via stratified random sampling to comprise a range of scores) based solely on the automated feedback. For example, some feedback suggested the essay might lack a clear thesis statement. In this case, the researchers added the missing essay feature. Other feedback might comment that the essay would benefit from more elaboration or examples. In these cases, one or two examples and sentences were added to each body paragraph.

For the complete corpus of 201 essays, W-Pal scores were significantly and positively correlated with human ratings ($r = .48, p < .001$), but tended to underestimate human scores by about 0.5 to 1.0 points. Revised essays were scored significantly higher than original drafts. This indicates that students who followed W-Pal feedback would potentially improve their essay scores. However, and more importantly, this process also revealed challenges to the effectiveness of the feedback system. First, we observed that the quantity of feedback given was sometimes overwhelming. Essays that surpassed the basic thresholds could trigger four or five feedback messages on introductions, bodies, conclusions, and overall revising. Second, as we submitted essays for usability testing, it became clear that the various versions of feedback messages were too repetitive. Over the course of several weeks or months, students had the opportunity to submit numerous drafts and revisions, often triggering the same feedback. Struggling writers who often make the same errors could grow frustrated with highly repetitive feedback.

The final issue identified was that the essay feedback necessarily addressed broad goals rather than specific errors. This feedback was due to the nature of the NLP algorithms implemented in that version of W-Pal. This lack of specificity was assessed within the internal usability process as potentially hindering students’ ability to utilize the feedback effectively. Without specific guidance, some students might find it difficult to implement strategy suggestions, although the ideal level of specificity is unknown. Consequently, in subsequent testing with end-users (i.e., students), these concerns were targeted explicitly via questionnaires and open-ended user comments on the system.

Overall, the development of W-Pal Version 1 was informed by a variety of usability assessments that guided the design of system components. However, W-Pal is intended to be used by students as a holistic system for learning and practicing writing strategies, which may occur in class or at home. Such learning occurs over several weeks, a semester, or a school year. To test the feasibility of W-Pal, it was necessary to deploy the system in an ecological setting over a longer span of time. This method of in vivo testing can help to identify how a system performs in closer connection to its intended context and environment (Wixon, 2003). That is, the purpose of this study was not to establish that W-Pal was effective, but to assess the feasibility of the system in a naturalistic setting—to identify potential threats and barriers to effective deployment that could be then addressed in the final stages of system development.

3. In vivo feasibility study of Writing Pal version 1

3.1. Participants

Two high school English teachers and 141 tenth-grade students participated in a feasibility study over six months (November, 2010 to May, 2011). The high school was located in the Washington, DC area and enrolled over 2,400 students: 49.0% female students, with 22.3% Asian, 4.2% Black, 9.0% Hispanic, and 59.9% White students. 7.0% of students were described as having limited English proficiency, and 10.9% qualified for free or reduced-price meals.

3.2. Measures

Students wrote timed, prompt-based essays on one of two counterbalanced topics at the beginning and end of the study. Students also completed several surveys embedded in the system regarding their perceptions of the lessons and completed two offline surveys regarding their perceptions of the practice games and automated feedback. These
measures included both multiple-choice ratings and open-ended questions. In this paper, we focus on students’ responses to the open-ended items, which helped to identify specific usability issues in W-Pal.

3.2.1. Pre- and post-study essays

Students wrote timed (25 minutes), prompt-based essays on two topics similar to SAT exam writing prompts. One writing prompt asked students to argue whether cooperation or competition leads to greater success; a second prompt asked students to argue whether or not images and impressions have too great of an effect on peoples’ decisions. Essays were written offline (i.e., not within W-Pal), transcribed by the researchers, and assessed via NLP algorithms utilizing Coh-Metrix and other text analysis tools (e.g., Crossley & McNamara, 2011; Graesser & McNamara, 2011). A holistic score was calculated using a combination of threshold and regression techniques including essay features such as length, keywords, lexical sophistication, structure, and cohesion (Crossley, Roscoe, Graesser, & McNamara, 2011). The reliability of this algorithm, based on a separate test set of 105 essays and expert human scores, was 39% perfect agreement (exact match of computer and human scores) and 92% adjacent agreement (scores within one point).

Student essays were also scored by a team of independent human raters using a six-point scale based upon the SAT scoring rubric. The raters were trained on the rubric until they reached an inter-rater reliability correlation of .70 or higher, and then two raters scored each essay. Human ratings that differed by one point were averaged (i.e., a score of 2 by rater 1 and a score of 3 by rater 2 would result in a final score of 2.5). If any scores differed by more than one point then a third judge adjudicated the difference. However, pairs of human raters assigned scores that were in exact agreement or were within one point more than 95% of the time.

3.2.2. Lesson perception survey

After viewing each lesson, a five-item survey appeared in a pop-up window. Using a four-point scale, students rated how much they learned and whether they would be willing view the lesson again. In three open-ended items, students described the “most helpful information” they learned, described their perceptions of the animated characters, and provided suggestions for “how to improve this lesson.”

3.2.3. Game perception survey

After interacting with W-Pal for several months, students completed a four-item feedback survey of their perceptions of the games. Using a four-point scale, students rated games regarding their helpfulness and enjoyability for practicing writing strategies. In two open-ended items, students made suggestions for improving the helpfulness of the games and redesigning the games to be more enjoyable.

3.2.4. Feedback perception survey

Along with the Game Perception Survey, students also completed an eight-item survey of their perceptions of the essay writing tools and feedback. Using a four-point scale, students rated the overall difficulty of using the essay writing interface, the difficulty of individual tools, the appropriateness of feedback quantity, the understandability of the feedback, and the usability of the feedback. Two open-ended items asked students to offer suggestions for making the feedback more “clear, more understandable, or more usable” and to suggest what “essay features or writing strategies” should be included in future feedback.

3.3. Procedure

Students wrote a pre-study essay in November. Throughout the school year, teachers and students incorporated W-Pal into the curriculum of their high school English classrooms. Students viewed the lessons and played the games, wrote several practice essays, wrote essays assigned by teachers, and completed the surveys. Interestingly, the teachers chose to maintain more strict control over students’ progress through the system—each component was tied to a specific assignment and free exploration was somewhat discouraged. Essays assigned by the teachers were often related to reading assignments, such as Molière’s Tartuffe. Students wrote a final post-study essay in June. As this was an ecological setting, some students did not complete all assignments; thus, reduced sample sizes are indicated when relevant.
3.4. Changes in essay quality

For students who wrote both a pre- and post-study essay (n = 113 students), scores assigned by NLP algorithms increased from an average of 2.3 points on a six-point scale prior to the study (with a standard deviation [SD] of 0.8) to 2.8 points (SD = 0.8) after the study. A paired-samples t-test was conducted to assess whether these gains were due to chance, which found that the gains were statistically reliable: \( t(112) = 5.85, p < .001 \). The effect size was \( d = .64 \), indicating a moderate effect. According to NLP assessments of textual features, post-study essays were also longer, showed clearer structure, used more concrete and precise wording, and displayed more lexical diversity. Post-study essays also showed more developed and elaborated content, with less repetition of themes and less hedging (e.g., words such as “maybe” or “might”).

Human essay ratings for the 113 students increased from an average of 3.0 points (SD = 0.6) for pre-study essays to 3.3 points (SD = 0.6) for post-study essays. This difference was statistically reliable, \( t(112) = 4.95, p < .001 \), and the effect size was moderate, \( d = .52 \). Human and automated ratings for pre-study essays were significantly and positively correlated (\( r = .52, p < .001 \)), as were human and automated ratings for post-study essays (\( r = .61, p < .001 \)). Thus, human and automated scores were relatively aligned, with human giving slightly higher scores.

To further convey the kinds of growth exhibited in students’ writing, we present examples from two students’ pre- and post-study essays along with their perceptions of the system (all responses and excerpts are presented as written without editing for spelling or other typos). Although gains cannot be directly attributed to the effects of W-Pal from this feasibility study, it is important to establish that students did improve over the course of their English class. Moreover, it is important to demonstrate that students viewed W-Pal as a beneficial and informative piece of their English composition curriculum.

The first examples are drawn from Harriet (a pseudonym) who wrote on “cooperation and competition” for her pre-study essay and on “images and impressions” for her post-study essay. Harriet’s pre-study essay received a “2” from both the human raters and W-Pal. This essay began with a clear thesis statement and presented several arguments to be discussed, but lacked strong evidence and support. Indeed, the body of the essay was a single paragraph without a clear topic sentence, factual information, or examples. Additionally, the concluding paragraph introduced new concepts that were not addressed previously and failed to summarize arguments. Thus, although this essay was on-topic and makes several points, the student appeared to lack a strong understanding of the appropriate organization and development of a persuasive essay:

People are taught to achieve success in different ways. Using cooperation, one can achieve more success because two hounds are better than one, it puts less emotional and physical stress one someone, and it means less work for you.

In your job or at school there is constant competition, such as in an attempt to get a raise or on a test or quiz. Sometimes the competition might not be against someone else, but it still puts an immense amount of stress on your body. However, cooperation is used more often. On teams of any sort we rely on one another to do their part to be victorious. This is cooperation. At times cooperating can be difficult, but when you succeed it is all worth it in the end.

Cooperation and competition have been a large part of our world, since Columbus sailed the ocean blue with the money provided by Spain’s monarchy. With this in hand cooperation can be useful in ones success. (Harriet, Pretest)

Throughout the school year, Harriet engaged with W-Pal by viewing the lessons, playing the practice games, and writing essays assigned by the instructor. In surveys distributed throughout the course, Harriet expressed specific benefits for various aspects of W-Pal, such as “Wpal helped my for my introductory paragraph,” and that she benefitted from “TAG and the examples of how to grab your readers attention.” In regards to organization, a problem with her pre-study essay, Harriet noted that “planning before an essay gets you a higher grade.” Additionally, Harriet found the mnemonic devices useful for remembering key goals, stating that “using mnemonics that stood out to me helped.”

Harriet’s post-study essay evidenced incremental gains in terms of writing quality and was scored a 3.5 by human raters and a 2.15 by W-Pal. As with her pre-study essay, the post-study essay offered a thesis, but now included a clearer preview of her arguments. Harriet also incorporated the “ask a question” attention-grabbing technique. The body of the posttest essay was divided into three distinct paragraphs that began to develop the claims previewed in
the introduction. Although the evidence was still underdeveloped, it incorporated more specific examples and ideas. Lastly, the conclusion of the essay still lacked a concise summary of main points but no longer introduced new material. Thus, Harriet’s post-study essay showed moderate improvements from her earlier writing, and some improvements map onto lessons and concepts she observed in her studies with W-Pal:

Can you imagine how your life would differ if you were blind? Color and beauty would have a whole new meaning. Images and impressions have too much of an effect on people because magazines and fashion don’t show true appearances, we judge people too quickly, and we don’t give people the benefit of the doubt. Prada, Gucci, and Coach. When we see these clothes on the runway or in a magazine, we see thin, tall, beautiful woman and our brain thinks that we too should resemble them. Our society is based on the outer image. At the pool a boy walks up to a girl wearing a bikini that fits her body well and passes the girl wearing a one piece to cover her belly. Maybe this was just her luck, but time and time again she becomes self conscious. Boys aren’t afraid to be judgmental before they ever speak to a person. I’m Not saying girls aren’t either, but they just have a better way of hiding it. Remember don’t judge a book by its cover.

Impressions affect our opinion on others and people will put them above all other instances. For example, if I am feeling sick and I snap at my mother while she is talking to her friend, her friend may think I’m a brat and avoid me at other events.

We forget that we should be the best that we can be. Treat others the way you would want to be treated. Also, don’t give an opinion without knowing the full story or reality. (Harriet, Posttest)

The second set of example essays was drawn from Leo (a pseudonym). Unlike Harriet, Leo’s pre-study essay exhibited a clearer structure, including an introduction that previewed some of the ideas to be developed in the body, and a clearly-marked conclusion that summarized major points. Thus, Leo appeared to grasp some of the structure and goals of a persuasive essay. However, the development of the evidence in Leo’s three body paragraphs relied heavily on unsupported hypothetical and opinion-based claims rather than fact-based claims. His communication of these ideas also lacked strong cohesive cues and transitions. This essay was given a score of 2.5 by human raters and 3.43 by W-Pal:

A wide range of people have differing oppositions on whether cooperation or competition is more effective. I believe cooperation is the key on to success, as two brains contain more knowledge than one brain. In fact, throughout history cooperation has led to good ideas and practices, and competition has led to struggles. Throughout time, competition has led to corrupt practices and teachings as well.

I believe cooperation is more successful than competition. When people compete, one person prevails and one fails. However with cooperation, all individuals prevail and succeed. In cooperation, people may not agree with other philosophies, however they get used to the idea. However competition leads to conflict and unneeded struggles. All in all cooperation, in my opinion I believe is easier to work with than competition.

Without competition, the make-up of the United States would be completely different. When people cooperate, people do not express true opinions always. This is good and as a result people can work together, without the rule of a riot. In competition, in the end noone wins because one individual succeeds but he also makes an enemy. Without cooperation, the world may look completely different. If the United States did not cooperate with allies or did not have allies, our country may be attacked at any time. by any countries without prior notifications, competition also increases number of enemies and decreases your allies, with competition individuals may think they “win” however in the long run they really lose.

In conclusion, I truely believe cooperation has led to unneeded struggles and cooperation has led to success. Without cooperation, the world would be in caos. Although a little competition is good, it can also lend to number of enemies. Cooperation is the key to success, cooperation helps build allies rather than decreasing allies. (Leo, Pretest)

Throughout the school year, Leo appeared highly-motivated and engaged in the W-Pal system. In his feedback, he wrote, “Thank you for helping me better enhance my SAT writing score! I believe I will be successful due to your W-Pal Lessons.” The student specifically mentioned freewriting as a favorite strategy: “FAST PACE is going to help me write better essays! I learned important acronyms, and information. I learned to think about the prompt, add questions, think about the opposing side, ect.” For most modules, the student expressed some personal benefit, such as learning “to support my ideas” from Body Building, and that “the acronym Tag really helps me understand how to write an
introduction” from Introduction Building. Overall, Leo’s enthusiasm for the system and willingness to improve seemed to make W-Pal a positive experience for him.

Leo’s posttest essay was given a score of 3.5 by human raters and 3.94 by W-Pal. As with his pre-study essay, the post-study essay followed a clear structure. The introduction contained a clear thesis, previewed his arguments, and included an attention-grabbing technique (historical example). Likewise, the conclusion provided a summary of the main points and connected to the reader using a recommendation. Compared to the pretest, Leo’s posttest essay perhaps showed the most growth with regards to his use of evidence. The student now drew upon somewhat more objective, real-world references from history, popular culture, personal experience, and literature read in his class. The posttest essay also incorporated more cohesive devices, such as transition and connective phrases that linked ideas (e.g., “for example” and “yet again”). Leo’s essay still displayed problems with unsupported claims and a narrative writing style (i.e., use of the first person perspective), but showed growth from the pre-study essay.

Images and impressions exert the evil force existent in the human body. To start off, images and impressions are inaccurate portrayals and views of a person in reality. For example, in Japan 1,000 years ago, the known rulers embodied a figure head whereas behind the scenes political employees were imposing strict regulations. Secondly, people are too caught up in the perfect idea of a celebrity. Furthermore, people ought to rid of impressions as they are only skewed views of reality. Ideas and impressions lead people to commit unideal acts, in addition to acting out of character.

To start off, in ancient times many rulers ruled as if they dominated all the land. However, often times, the absolute political ruler often procured most of the fame although he merely didn’t influence political structure, as his advisors took up on the majority of the work. Likewise, present day media such as the People magazine inaccurately portrays the lives of rich and well-off celebrities in an attempt to enhance their image. While impressions play a influential role in our society I believe they are false and irrelevant to a persons true character. Secondly, while I was a 12 year old boy meandering around at a Washington Wizards game, I encountered my hero or should I say figure hero. I finally had discovered the true character of Lebron James without mass-media bias. To start off, when I asked Mr. James for a simple photo he declined in a disdain and abrupt matter, I thought nothing of it than I asked for an autograph, yet again he shunned me. This very day I learned that the media inaccurately portrayed the character of my once greatest hero, Lebron James.

Lastly, in the book Antigone, Antigone is portrayed as a weak, and rather unintelligent being. Yet through her actions and conversation with Creon it is revealed that she has a very strong wiled and rigid nature. She wouldn’t take no for an answer. Yet again, my first impression of yet another person has changed.

All in all, images and impressions are not relevant to a character of a person. Images and impressions are human nature; however, they reveal nothing of a persons personality or traits. If one must discover the true character of a person they must eliminate the influential role of images and impressions in the society in which we live in today. (Leo, Posttest)

In sum, the above excerpts help to exemplify the type of writing improvements students experienced in a tenth-grade English course that incorporated W-Pal. To be clear, students did not become world-class writers in just one school year. However, noticeable gains were observed in their writing proficiency, some of which paralleled self-reported perceptions of W-Pal. Thus, as stated earlier, we cannot make strong causal claims about the efficacy of W-Pal from this feasibility study, but such findings do suggest that W-Pal is a plausible and potentially beneficial tool to explore in high school composition instruction. In the remaining sections, we consider students’ self-reported perceptions of the system in more detail.

3.5. Students’ perceptions of the lessons

Overall, students perceived the lessons to be helpful and informative. Over half of the students reported learning three or more “new ideas” from W-Pal. When asked to describe the most helpful ideas learned in the lessons, students most frequently cited the mnemonic devices, although other strategies were also specifically mentioned. Many students stated that the strategies learned in the lessons helped them to better understand and enact the writing process. Although students found the lessons beneficial, they also disapproved of the presentation style of the lessons. In particular, students disliked the animated characters, frequently describing them as “awkward” and “boring.” Students also critiqued the dialog between the agents, and requested straightforward instruction with more competent-seeming characters. Such
comments are nicely exemplified by one student’s feedback, who wrote, “Their voices are very robotic and the lesson was way too long. Maybe if it was split into several sections then it would be easier to concentrate on the task.” Overall, both students and teachers requested that the lessons be shorter and faster, while retaining all of the information.

3.6. Students’ Perceptions of the Practice Games

Across sampled games, the large majority (80.1% of students) rated the games as somewhat helpful or very helpful for practicing the writing strategies covered in the modules. Similarly, well over half of the students (65.5%) rated the games as somewhat enjoyable or very enjoyable to play. Most students felt that the games they played were beneficial and engaging. However, responses to open-ended questions also highlighted several ways in which students considered that the games could be improved. First, students requested that the games be more generative and interactive, rather than presenting identification and recognition tasks. For example, one student commented, “I think there should be a challenge where we actually use the strategy instead of finding them in essays.” Thus, students desired more difficult forms of practice where they could actively apply the writing strategies to their own work. Other students encountered problems with the clarity of the game instructions, becoming confused about the nature of the learning task or game mechanics. Finally, some students noted that the graphics and sound effects in the games needed to be improved, or they commented that some games were “not very fun because they had a learning element that was very obvious” and needed “more pictures and music and less words.” In sum, students reported that the games were helpful and enjoyable, but these perceptions may have been negatively impacted by games that lacked sufficient challenge, lacked interactive game-play, or had unclear directions.

3.7. Students’ Perceptions of the Essay Writing and Feedback Tools

Overall, most students (81.5%) rated the essay writing tools as easy or very easy to use. However, two specific features frustrated some students: 23.7% of students considered reading the feedback somewhat difficult or very difficult, and 24.6% of students considered revising their essays somewhat difficult or very difficult. This may have been due to the quantity and clarity of the feedback. Although half of students reported that they received an appropriate amount of feedback (49.5%), others reported that they received not enough (38.8%) or too much (11.6%). Thus, these results confirmed the observations from our internal usability testing (Roscoe et al., 2011), which suggested that the highly variable amount of feedback given could be problematic. Similarly, although most students rated the feedback as understandable (61.2%), about one-third of students rated the feedback as somewhat confusing or very confusing (38.3%). Despite these challenges, most students (78.6%) reported that the feedback was useful overall.

Students’ open-ended responses revealed their concerns more directly. Specificity was a common issue; students requested more detailed feedback that was specific to their current essays, feedback that provided explicit examples on how to improve their writing, and feedback that addressed particular essay features (e.g., topic sentences). For instance, one student wrote, “The feedback needs to be more helpful for us on our own personal essay. Not just general feedback. I don’t know what I did wrong in my essay.” Other students requested that the feedback system provide information on both the positive and negative aspects of their essays (i.e., both praise and critical comments).

4. Discussion

W-Pal is designed to provide strategy instruction that spans multiple phases of the writing process, offers extended game-based and whole-essay practice, and provide automated formative feedback and scoring. A multiple-method approach to usability testing has been an essential component of the system’s development. The combination of focus groups, experiments, internal testing, and in vivo testing allowed us to capitalize on the strengths of each method while more easily addressing their weaknesses (Holzinger, 2005).

Traditional usability methods, such as focus groups and small experiments, recruit relatively small numbers of end-users to rapidly gather data on the functionality or efficacy of specific components of the system. System errors and obstacles can be quickly addressed over multiple cycles of testing until necessary modifications become minor or inconsequential. In W-Pal development, focus groups with teachers informed the core content to be taught by the system. From teachers, we learned that W-Pal would need to be a comprehensive and flexible system that could fit smoothly into diverse writing curricula. Subsequently, prototype versions of the writing strategy modules were developed and tested
via iterative component studies. These studies revealed the need for more interactive and engaging lessons and practice, which prompted the development of numerous checkpoint and practice game activities. Yet another usability testing format was needed to evaluate our feedback system prior to deployment with students. In this case, internal usability testing is well-suited to assessing whether a system is working as intended and for identifying potential problems that might negatively impact end-users without putting those users at risk. This testing not only indicated that the feedback system provided useful information that could be used to effectively revise essays, but also revealed that caution might be needed regarding feedback specificity and quantity.

Based on the results of usability testing in the early and middle phases of W-Pal development, a complete Version 1 system was constructed. As with most ITSs, W-Pal was a complex system with many components, and was designed to be used as a holistic system in authentic classrooms. Although lab-based studies focusing on individual system components were highly informative, they provided an incomplete picture of the feasibility of the system. Thus, in vivo testing was necessary to study how interactions among system components, over educationally-meaningful lengths of time, influenced user perceptions and system functionality. Although much slower than other usability methods, in vivo testing provided rich data through which end-users could evaluate the complete ITS in its intended environment (Wixon, 2003).

W-Pal was deployed with over 140 high school English students. Most aspects of W-Pal were judged as beneficial sources of writing strategy information, practice, and feedback. Students could articulate specific strategies and mnemonics learned from the lessons and games, and rated these tools and essay feedback as helpful and easy to use. Nevertheless, students identified design issues that could be remedied to enhance their W-Pal experience, which may, in turn, improve system utility. First, students argued that the lessons were too long and didactic and disliked the cartoonish animated characters and their dialog. Second, students expressed interest in more difficult games that required active text generation and commented that the games could be improved via the addition of more “game-like” graphics, sounds, and features. Finally, students wanted feedback that addressed specific strengths and weaknesses of their essays and a perceived lack of specificity undermined some students’ confidence in the system.

4.1. Development of Writing Pal Version 2

In response to the results of our multiple-method usability and feasibility testing, many changes were made to W-Pal. Version 2 comprised an overhauled series of strategy lessons, several new practice games, and a new essay feedback interface. These modifications are described in the next sections along with considerations for future research and classroom integration.

4.1.1. Revised lesson videos

The strategy modules were retained, but with a significantly altered structure and format. Strategy lessons were divided into series of 38 five-minute, strategy-focused videos (Table 1). Each module begins with an Overview that presents the rationale and preview of the module (see Figure 4). Specific strategies are then discussed in 3 or 4 strategy videos. Many mnemonics were retained, such as TAG and RECAP. Other mnemonics were revised to facilitate recall. For example, the FAST PACE mnemonic was shortened to FAST (i.e., Figure out the prompt, Ask and answer questions, Support arguments with evidence, and Think about the other side). This reformating addressed concerns about length, boredom, and distracting dialog; Version 2 lessons were much more straightforward and informative. Mike and Sheila’s scripts were edited to make them knowledgeable and competent, and the teacher was changed to Mr. Evans to offer a more appealing TTS voice. Each video was presented by a single character instead of multiple agents who engaged in extraneous dialog.

Checkpoints were moved to the end of each video and standardized with two formats: a quiz and a game show. Students could select which version they would like to complete. Both formats presented multiple-choice questions tapping students’ recall and application of the strategies. In the game show, two opponent characters were shown who also

---

1 After the acceptance of this manuscript for publication, the authors conducted an experimental evaluation of the W-Pal Version 2 system. Over 85 high school students from an urban area in the southwest United States participated in a 10-session program based on W-Pal. W-Pal users improved their overall writing proficiency and cohesion, improved their writing strategy knowledge, and demonstrated more substantive essay revising (e.g., Crossley, Varner, Roscoe, & McNamara, 2013; Roscoe, Brandon, Snow, & McNamara, 2013; Roscoe, Snow, & McNamara, 2013; Roscoe, Snow, Varner, & McNamara, in press).
Table 1

<table>
<thead>
<tr>
<th>Module</th>
<th>Content</th>
<th>Strategy Lesson Videos&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prologue</td>
<td>Introduces W-Pal, the animated characters, and the value of writing</td>
<td>Meet the Students; Practice Makes Perfect</td>
</tr>
<tr>
<td>Freewriting</td>
<td>FAST mnemonic for quickly generating essay ideas, arguments, and evidence prior to writing</td>
<td>Figure Out the Prompt; Ask and Answer Questions; Support with Evidence, and Think about the Other Side</td>
</tr>
<tr>
<td>Planning</td>
<td>Formal and informal strategies for organizing arguments and evidence in a clear essay structure</td>
<td>Positions, Arguments, and Evidence; Outlines; and Flowcharts</td>
</tr>
<tr>
<td>Introduction Building</td>
<td>TAG mnemonic for writing effective introduction paragraphs</td>
<td>Thesis Statements, Argument Previews, and Grab the Reader’s Attention</td>
</tr>
<tr>
<td>Body Building</td>
<td>CASE mnemonic for writing body paragraphs that present concise arguments with supporting evidence</td>
<td>Topic Sentences; Evidence Sentences; and Strengthening Your Evidence</td>
</tr>
<tr>
<td>Conclusion Building</td>
<td>RECAP mnemonic for writing conclusion paragraphs that restate the thesis and summarize main ideas, without adding new evidence</td>
<td>Summarize the Essay; Close the Essay, and Hold the Reader’s Attention</td>
</tr>
<tr>
<td>Paraphrasing</td>
<td>Strategies for expressing ideas using more varied and precise wording and sentence structure</td>
<td>Synonym Strategy; Structure Strategy; Condensing Strategy; and Splitting Strategy</td>
</tr>
<tr>
<td>Cohesion Building</td>
<td>Strategies for adding cohesive cues to text and improve readability</td>
<td>Signpost Strategy; Threading; and Connectives Strategy</td>
</tr>
<tr>
<td>Revising</td>
<td>ARMS mnemonic to provide a variety of ways to edit and revise essay drafts</td>
<td>Add More; Removing Irrelevant Details; Moving Essay Sections; and Substituting Ideas</td>
</tr>
</tbody>
</table>

<sup>a</sup> All modules also contain an overview video that introduces the topic and strategies.

Figure 4. Version 2 Body Building lesson is divided into four short lessons that introduce the CASE mnemonic (top-left), offer strategies for topic sentences (top-right), discuss types of evidence (bottom-left), and tips for improving evidence (bottom-right).
earn points for correct answers. To win, students must answer more questions correctly than the computer opponents. In addition, a Module Map interface was added to accommodate the new lesson structure and show progress through the lessons and games. Importantly, this system was tiered; students “unlocked” content by completing prerequisites. Completion of the overview unlocked specific strategy lessons, which in turn unlocked relevant educational games to practice those strategies.

4.1.2. Revised practice games

Several games were updated to improve graphics and functionality. For example, CON-Artist helped students practice the Connectives Strategy (Cohesion Building). Players tracked a thief by following a series of clues, which took the form of sentences missing a connective phrase. This game was updated with new graphics and the ability to play an easy, medium, or hard version. Similarly, in Version 1, students practiced freewriting by playing Freewrite Feud. In this game, students freewrote on a prompt and earned points by writing quickly and incorporating hidden prompt-based keywords. Usability testing revealed that students had difficulty triggering more than two hidden words, but could inflate their scores by typing rapidly with minimal idea generation. In Version 2, this game was replaced with Freewrite Flash (Figure 5). The task remained similar, but keyword lists and scoring algorithms were updated to be more sensitive to ideas contained within the freewrites. Game instructions and graphics were also changed to further emphasize idea generation.

Three new games were added to further address students’ concerns about interactivity and difficulty: Planning Passage (Planning Module), RoBoCo (Body Building Module), and Lockdown (Conclusion Building Module) (see Figure 6). For example, W-Pal Version 1 did not offer generative practice for Body Building. A new game, RoBoCo, was added to enable such practice. Players took on the role of a robot designer who must build robots to display at an annual show. Players earned robot heads and bodies by writing topic sentences and supporting evidence. Next, players built robots using a drag-and-drop interface. Players earned more money (i.e., a higher score) based on the quality of their work and the number of robots built. Thus, this game challenged students to generate text rather than recognize examples and was more engaging and interactive by allowing students to build customized robots.

4.1.3. Revised essay feedback reports

The Essay Writing Interface was largely retained. However, because some students encountered problems with feedback quantity and specificity, the reports were significantly updated. Students still received a holistic rating (from “Poor” to “Great” on a 6-point scale). Feedback recommendations were rewritten to ensure that problem-identification was stated in an impersonal and suggestive manner, rather than personal and definitive. Instead of stating “Your essays lacks...” the feedback stated “This essay may lack...” In contrast, strategy suggestions for problem-resolution were rewritten to be both personal and specific, such as “One way that you can organize your ideas...” The purpose of these changes was to a) reduce any sense of threat from receiving critiques from an automated system and b) empower students to feel capable of strategically revising their work. More accurate essay scoring algorithms were also developed to improve scoring and feedback validity.

Two new functions were added to offer students more control over feedback quantity and content (Figure 7). Students could now use a More on this Topic button to request additional feedback on the same topic as the original message. In addition, students could now use a More on a Different Topic button to request feedback on an alternate topic. In this case, students received feedback on the next level up in the threshold hierarchy (i.e., Legitimacy, Length, Relevance, Structure, Introductions, Bodies, Conclusions, and Revising).

4.2. Future research and potential for classroom integration

The functionality and design of W-Pal differs significantly from many existing AWE tools. From the outset, W-Pal focused on strategy instruction rather than essay scoring, which has informed the development of a modular and flexible system with many components. Key questions for practice and future research relate to how W-Pal can be effectively integrated within a curriculum and what specific game and feedback features promote effective practice.

First, the modular structure of W-Pal affords a variety of curricular decisions. One way to integrate W-Pal may be to assign particular modules in advance of classroom discussions. That is, W-Pal can be used as a preparatory activity to stimulate thoughtful discussions of writing principles in class. Alternatively, teachers can provide core instruction on writing principles prior to assigning W-Pal for homework. The strategies taught in W-Pal can bolster the repertoire
of strategies presented in class. Such reinforcement can also be targeted; students can be assigned to work on content that addresses their individual writing needs.

Given that W-Pal instruction can be structured in different ways (e.g., more or less freedom to access system features), what level of choice is optimal? Research suggests that academic task choices can increase student motivation (Patall, Cooper, & Robinson, 2008), and students who expend more effort practicing a strategy should acquire that strategy more readily. However, prior skill may influence these effects. In open-ended learning tasks and environments, such as web-search and hypermedia, students with low prior knowledge often struggle to navigate the system (McNamara & Shapiro, 2005; Scheiter & Gerjets, 2007). Thus, in W-Pal and other systems that provide students with significant instructional freedom for writing, it is essential to understand if the level of freedom provided will need to be limited for
less-skilled students. That is, research is needed to better understand when instructors should carefully guide students through the use of W-Pal in a curriculum or when a hands-off approach is warranted.

W-Pal offers similarly diverse options for writing practice. The games enable practice of specific strategies and goals, whereas the essay writing tools allow practice of strategies in the context of whole essays. Such opportunities can be sequenced in different ways. For instance, students might write complete essays first, as a way to diagnose writing problems, and then play practice games that target their individual needs as revealed by feedback. Alternatively, students might initially practice a variety of strategies using the games, and then write essays to practice synthesizing
and integrating individual strategies. The design of the essay feedback also grants teachers and students some control over the feedback students receive. Thus, teachers can assign students to focus on specific aspects of revision or on more global revisions of essays.

Second, what game features are most supportive of motivation? Gaming environments differ widely, ranging from accruing points for completing simple objectives to full-fledged narratives with complex rules and rewards (Gee, 2007). Well-designed games are appealing because they tap positive affective states and expectations (Neil, Wainess, & Baker, 2005). However, the framing of the game and content is critical. Games can be used to capture student interest, but designers must take care not to detract from pedagogical and educational aims (Gredler, 2004). Future studies on W-Pal, and systems that use educational games for writing, will need to explore what game elements are supportive of writing practice. Our data suggested that such games need to incorporate generative writing opportunities, but it is not clear whether features such as narrative and competition should be emphasized or de-emphasized.

Third, another key aspect of strategy instruction is formative feedback on students’ learning and progress (Shute, 2008), which is considered a key component in feedback on student writing (McGarel & Verbeem, 2007; Sommers, 1982). Although published work has assessed scoring accuracy, relatively few studies have examined the effects of AWE feedback on writing proficiency. Yigal Attali and Jill Burstein (2006) found that students using ETS’s Criterion system wrote final drafts that earned higher scores (d = .47, moderate effect), were longer, contained fewer errors, improved in stating and supporting main ideas, and were more likely to contain a conclusion. However, no control condition was available, and so gains cannot be fully attributed to use of Criterion. JoAnn L. Rock (2007) and Ronald T. Kellogg, Alison Whiteford, and Thomas Quinlan (2010) used experimental designs to assess the efficacy of Criterion versus instructor feedback. Both of these studies observed that students’ essays improved regardless of whether AWE feedback was provided, although receiving AWE feedback seemed to reduce superficial errors of grammar, mechanics, usage, and style. In sum, available evidence suggested that AWE feedback could improve student writing, but more research is needed to explore the conditions that determine how and when that feedback is most effective.

What forms of writing feedback are most appropriate and effective in computer-based tools for writing? An important distinction between W-Pal and AWEs is that W-Pal focuses on strategy instruction and formative feedback instead of feedback on style, mechanics, or grammar. Moreover, a common approach in existing automated writing systems is to provide an extensive, multi-page report on numerous essay traits (Grimes & Warschauer, 2010). It is unclear how students respond to and comprehend differing types and amounts of feedback. Students who receive more feedback...
overall should produce better final drafts than students who receive less feedback because they will be exposed to more strategies for doing so. However, these effects could be moderated by students’ prior ability. For less-skilled readers and writers, receiving too much feedback may be overwhelming. Such students may be more likely to have difficulty reading and comprehending copious amounts of feedback and may have trouble implementing and balancing multiple demands for revising. Indeed, Douglas Grimes and Mark Warschauer (2010) reported that teachers often had to summarize and simplify feedback from an AWE system in order to make it usable and accessible for their students. Future research will need to consider whether a comprehensive or scaffolded approach to automated feedback is more efficacious.

Finally, another question may be the grade level and English language proficiency of W-Pal users. Our system was developed with English-speaking high school students in mind, but the strategies included in W-Pal may be relevant to advanced middle school students, struggling and remedial college-age writers, and English-language learners. To the extent that college freshmen lack a strong grasp of fundamental writing skills and strategies, tools like W-Pal may provide a means for reviewing and strengthening these basic principles. Similarly, W-Pal can help to make the often implicit “rules of writing” more explicit for those who are still improving their English fluency. Lastly, as this paper has discussed, there are numerous questions still to be answered in terms of the development of computer-based tools for writing. For pre-service teachers learning how to teach writing, W-Pal might serve as a useful starting place for valuable discussion, creativity, and critique regarding computers in the composition curriculum.

5. Conclusion

The Writing Pal provides a novel and useful technology to support writing instruction, and the current version of the system is grounded in years of careful design and testing. As the system transitions from a development phase to broader deployment, it is poised to enable exciting new contributions to the writing curriculum, and inspire innovative research on the affordances and limitations of digital tools for writing and language assessment and pedagogy.

Rod D. Roscoe is an Assistant Professor of Human Systems Engineering in the Ira A. Fulton Schools of Engineering at Arizona State University. His research examines the interaction of metacognitive, cognitive, and motivational processes in learning, and how these processes can be effectively facilitated via educational technology and games, strategy instruction, and peer support.

Laura K. Allen is a PhD student in Psychology and the Learning Sciences Institute at Arizona State University. Her research examines the cognitive processes and abilities underlying reading and writing proficiency, and also considers the impact of these factors in second language learning.

Jennifer L. Weston is a PhD student in Psychology and the Learning Sciences Institute at Arizona State University. Her research focuses on the psychometrics of learning and strategy development, and in particular, the role of individual differences in predicting student learning.

Scott A. Crossley is an Associate Professor of Applied Linguistics and ESL at Georgia State University. His research focuses on corpus linguistics and the application of computational tools in second language learning and text comprehensibility.

Danielle S. McNamara is a Senior Research Scientist at the Learning Sciences Institute and a Professor of Psychology at Arizona State University. The overarching theme of her research is to better understand cognitive processes involved in comprehension, writing, knowledge acquisition, motivation, and memory, and to apply that understanding to educational practice by developing and testing educational technologies.

References


McNamara, Danielle S., Jackson, G. Tanner, & Graesser, Art. (2010). Intelligent tutoring and games (ITaG). In Young K. Baek (Ed.), *Gaming for classroom-based learning: Digital role-playing as a motivator of study* (pp. 44–45). Hershey, PA: IGI Global.


